DS-GA 3001: Introduction to causal inference for data scientists

Course description: Causal inference is the science of analyzing causal relationships between events. What is the impact of advertising on demand? By how much will a graduate degree increase one's salary? What is the impact of minimum wage on wages and unemployment? These are questions that require the understanding of causal connections between a decision and its consequences. Surprisingly enough, causal inference tools have only been relatively recently developed and taken to data. We will present Rubin's model of potential outcomes, which will be our primary framework in the course. We will study cases when the analyst has the power to design a randomized experiment (Randomized control trials, field experiments, AB testing) and cases when the analyst is present with a dataset where randomization has not explicitly occurred (observational studies). Our course will close with an overview of special topics in causal inference, such as causal inference in networks, and successful examples of machine learning for causal inference.

Instructor: Michael Gill (NYU CDS), **mzgill@nyu.edu**, Office hours: Tuesdays, 3pm-4:30pm, CDS, Room 620

Teaching assistant: Lei Xu (NYU CDS), **lx557@nyu.edu**, Office hours: Wednesdays, 1:30pm-3pm, CDS, Room 665

Schedule:
Lecture: Tuesdays, 1pm-2:40pm, 60 Fifth Avenue, Room C12
Lab: Wednesdays, 8:35pm-9:25pm, 60 Fifth Avenue, Room C10

Prerequisites: A first course in probability and statistics, and be acquainted with linear algebra.

Evaluation: Class participation/attendance (10%), Homework (35%), Midterm (25%), Final Exam (30%).

Textbooks:
There is NO required textbook. However, recommended books include:
[IR] Imbens, G. and Rubin, D. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences. An Introduction*. Cambridge University Press.
[MW] Morgan, S., and Winship, C. (2014). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, 2nd Edition. Cambridge.
[PGJ] Pearl, J., Glymour, M., and Jewell, N. (2016). *Causal Inference in Statistics: A Primer*. Wiley.
[IW] Imbens, G. and Woodridge, J. (2007). *What's new in Econometrics*, NBER lectures.
[HR] Hernan, M. and Robins, J. (2017). *Causal Inference*.

Software: The core of the course will be taught using the statistical computing environment R, freely available from https://www.r-project.org/. If students prefer to use Python instead, that is fine, but students will be expected to submit working Jupyter notebooks on problem sets. In the first lab session we will help students install R Studio.

Datasets: will be available on the course webpage. Datasets used in lecture and lab include, but are not limited to:
- Lalonde training program evaluation data
- Bertrand-Mullainathan labor market discrimination data
- Angrist data on draft Vietnam lottery

Homework: There will be 5 problem sets throughout the semester, released every two weeks. They are not meant to be long or onerous, but to help clarify concepts relayed in lecture and lab. Students will be expected to submit problem sets as PDFs, and (when appropriate) share replication code alongside the writeup. Students that fail to submit both write-ups and replication code will not receive full marks on their homework.

Homework policies:
- *Late assignments* will receive a 10% deduction for each hour late submitted after the posted deadline, rounding up. For example, an assignment submitted 45 minutes late will receive a 10% deduction, while an assignment submitted 70 minutes late will receive a 20% deduction. However, there will be a 10 minute "grace period" before an assignment is considered late. Requests for homework extensions outside of these terms will not be granted. Please contact the teaching staff for any additional questions relating to this policy.
- *Student collaboration* on problem sets is not permitted. It is okay to ask each other clarifying questions about concepts relevant to the homework, but students should never explicitly share their work or answers with others.

Exams: There will be two exams throughout the semester: a *midterm* on 3/6/2018, and a *final* on 5/15/2018. Both will be closed notes, and no calculators allowed. Please note the timing of the final exam may change, but the midterm time is fixed. The final exam date/time is tentative estimate from the NYU Registrar (as of Jan. 17, 2018).

Part I: Introduction

LECTURE 1. CAUSAL INFERENCE: SOME MOTIVATING EXAMPLES. *1/23/2018*

- MW, Chapter 1
- Holland, P. (1986). "Statistics and Causal Inference." *Journal of the American Statistical Association.*
- Angrist, J. and Pischke, J-S. (2010). "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics," *Journal of Economic Perspectives*.

LECTURE 2. THE NEYMAN-RUBIN MODEL OF POTENTIAL OUTCOMES. *1/30/2018*

- MW, Chapter 2.
- IR, Chapter 1.
- Heckman, J. (2005). "The scientific model of causality." *Sociological Methodology*.
- HR, Chapter 1.
    - ***Homework 1 announced*** *in lab, due by 8:30 pm on 2/7/2018.*

Part II. Causal inference in an experimental setting

LECTURE 3. RCTs, AB TESTING, BUSINESS EXPERIMENTS (1). *2/6/2018*

- IR, Chapter 4.
- Imbens, G., and Athey, S. (2016). "The Econometrics of randomized experiments" in the *Handbook of Field Experiments*.
- List, J., and Rasul, I. (2011). "Field Experiments in Labor Economics" Chapter 2 in *Handbook of Labor Economics* Volume 4a, O. Ashenfelter and D. Card (editors), Elsevier.
- Deaton, A. (2009). "Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development" *Proceedings of the British Academy*.
    - ***Homework 1 due*** *by 8:30 pm on 2/7/2018.*

LECTURE 4. RCTs, AB TESTING, BUSINESS EXPERIMENTS (2). *2/13/2018*

- Bertrand, M.; Karlan, D., Mullainathan, S., Shafir E., and Zinman, J. (2012). "What's advertising content worth? Evidence from a consumer credit marketing field experiment" *Quarterly Journal of Economics*.
- Anderson E., and Simester, D. (2011). "A Step-By-Step Guide to Smart Business Experiments", *Harvard Business Review*.

LECTURE 5. NONCOMPLIANCE & INSTRUMENTAL VARIABLES (1). *2/20/2018*

- Angrist, J. D., Imbens, G.W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434): 444–455.
- MW, Chapter 9
- IR, Chapters 23, 24, 25.
    - ***Homework 2 announced*** *in lab, due by 8:30 pm on 2/28/2018.*

Part III. Causal inference in observational studies

LECTURE 6. INSTRUMENTAL VARIABLES (2) & INTRO TO OBSERVATIONAL STUDIES. *2/27/2018*

- Imbens, G. (2010). "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)", *Journal of Economic Literature*, 399-423.
- Imbens, G. and Rubin, D. (1997.) "Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance." *The Annals of Statistics* 25(1):pp. 305–327.
- MW, Chapter 5
- IR, Chapters 12-13.
    - ***Homework 2 due*** *by 8:30 pm on 2/28/2018.*

Midterm Exam: Tuesday, March 6, 2018. In class.

LECTURE 7. MATCHING ESTIMATORS. *3/20/2018*

- Stuart, E. (2010). "Matching Methods for Causal Inference: A Review and a Look Forward". Statistical Science.
- Ho, D., Imai, K., King, G., Stuart, E. (2007). "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." Political Analysis.
- Rubin, D., and Waterman, R. (2006). "Estimating the Causal Effects of Marketing Interventions Using Propensity Score Methodology." *Statistical Science*.

LECTURE 8. DIFFERENCES-IN-DIFFERENCES, REGRESSION DISCONTINUITY. *3/27/2018*

- MW, Chapters 6-7
- Lee, D., and Lemieux, T. (2010). "Regression discontinuity designs in economics" *Journal of Economic Literature*.
- Calonico, S., Cattaneo, M., and Titiunik, R. (2014). "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs." *Econometrica*.
    - ***Homework 3 announced** in lab, due by 8:30 pm on 4/4/2018.*

LECTURE 9. EXTENDING DIFFERENCES-IN-DIFFERENCES. *4/3/2018*

- Athey, S. and Imbens, G. (2006) "Identification and Inference in Non-Linear Difference-in-Differences Models," *Econometrica*.
- Abadie, A., Diamond, A., and Hainmueller, J. (2010) "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program."
    - ***Homework 3 due** by 8:30 pm on 4/4/2018.*

LECTURE 10. HIGH-DIMENSIONAL MODELS. *4/10/2018*

- Belloni, A., Chernozhukov, V., Fernandez-Val, I., and Hansen, C. (2017). "Program Evaluation and Causal Inference with High-Dimensional Data." *Econometrica.*
- Brodersen, K., Galluser, F., Koehler, J., Remy, N., and Scott, S. (2015). "Inferring causal impact using Bayesian structural time series models." *Annals of Applied Statistics.*
- Varian, H. (2016). "Causal inference in economics and marketing," *PNAS*.

Part IV. Special topics

LECTURE 11. PRACTICAL CHALLENGES WITH INFERENCE. *4/17/2018*

- Abadie, A., and Imbens, G. (2008). "On the Failure of the Bootstrap for Matching Estimators." *Econometrica*.
- Abadie, A., Athey, S., Imbens, G., and Wooldridge, J. (2017). "Sampling-based vs. Design-based Uncertainty in Regression Analysis." *Working Paper.*
- Young, A. (2017). "Consistency without Inference: Instrumental Variables in Practical Application." *Working Paper.*
- Ding, P., and VanderWeele, T. (2016). "Sensitivity Analysis Without Assumptions." *Epedemiology*.
    - ***Homework 4 announced** in lab, due by 8:30 pm on 4/25/2018.*

LECTURE 12. CAUSAL INFERENCE IN NETWORKS. 4*/24/2018*

- Jones, J., Bond, R., Bakshy, E., Eckles, D., and Fowler, J. (2017). "Social influence and political mobilization: Further evidence from a randomized experiment in the 2012 U.S. presidential election." *PLOS One*.
- Athey, S., Eckles, D., and Imbens, G. (2017). "Exact p-Values for Network Interference". *Journal of the American Statistical Association*.
- Eckles, D., Karrer, B., and Ugandar, J. (2017). "Design and Analysis of Experiments in Networks: Reducing Bias from Interference." *Journal of Causal Inference*.
    - ***Homework 4 due** by 8:30 pm on 4/25/2018.*

LECTURE 13. MACHINE LEARNING AND CAUSAL INFERENCE. *5/1/2018*

- Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. (2017). "Deep IV: A Flexible Approach for Counterfactual Prediction," *Proceedings of the 34th International Conference on Machine Learning, ICML 2017.*
- Wager, S., and Athey, S. (2017). "Estimation and Inference of Heterogenous Treatment Effects using Random Forests."
    - ***Homework 5 announced** in lab, due by 8:30 pm on 5/8/2018.*

Reading Week: No class, but Homework 5 due by 8:30 pm on 5/8/2018

Final Exam: Tuesday, May 15 from 2-3:50pm. Date and timing subject to change, from NYU Registrar